

Geuvadis eQTL methods

Tuuli Lappalainen

August 30, 2012

Data

- EUR = CEU+GBR+FIN+TSI, n = 373
- YRI, n = 89
- Genotypes
 - 421 samples with Phase 1 genotypes
 - 41 samples with imputed genotypes (Omni 2.5M -> Phase1)
- Expression
 - Standard eQTL analysis with exon quantifications
 - Test eQTL discoveries with transcript quantifications
 - Further analysis with splicing quantifications etc – see last slide

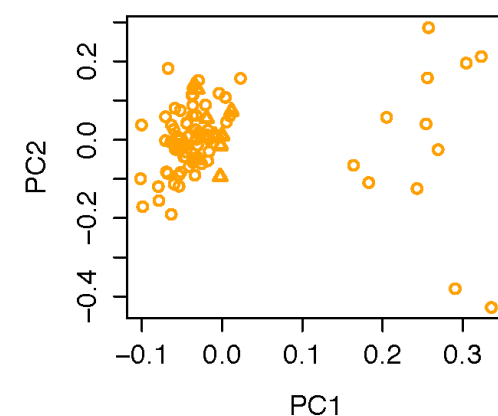
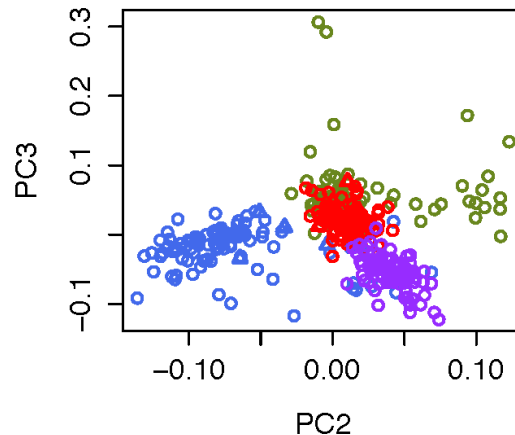
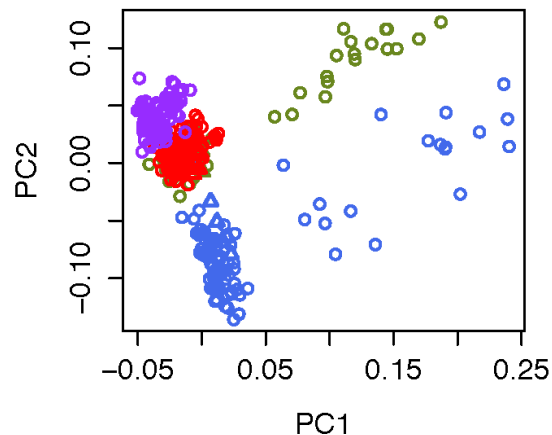
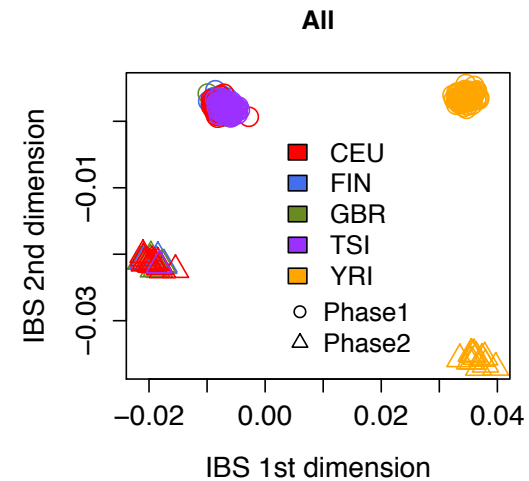
Genotype data covariates

Correction needed for two things:

1. Imputed *versus* Phase 1 samples
 - allele frequencies have been checked to be well correlated, but tiny differences across the genome add up to clear clustering of individuals by imputation status in IBS
 - this is normal and expected, and not a big concern
2. Other genotype batch effects
 - PCA shows some outliers, which correlates with some 1000g sequencing batches
3. Population structure
 - Analyze EUR and YRI separately
 - The European sample has to be corrected for genome-wide population structure
 - No admixed samples = no need for local ancestry correction

➤ **Correction: genotype covariates in the eQTL analysis**

- EUR: PC 1,2,3 + imputation status
- YRI: PC 1,2 + imputation status

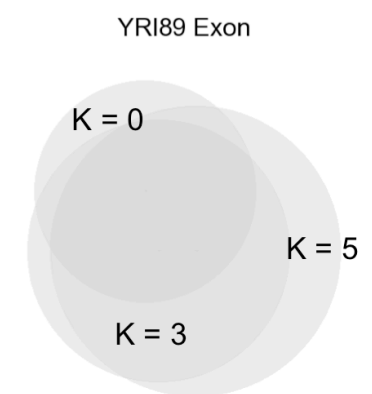
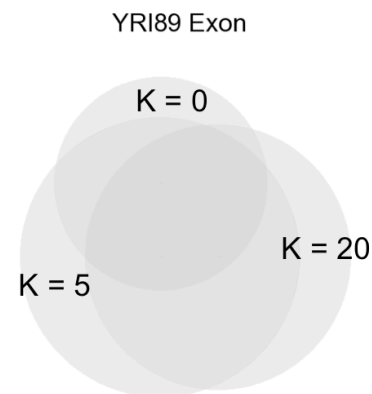
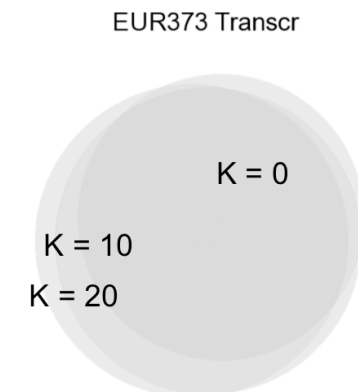
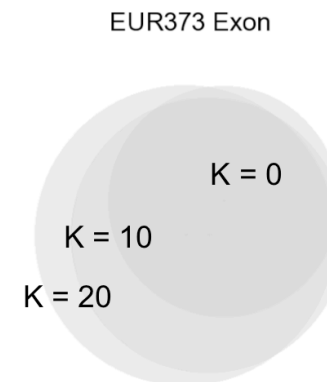
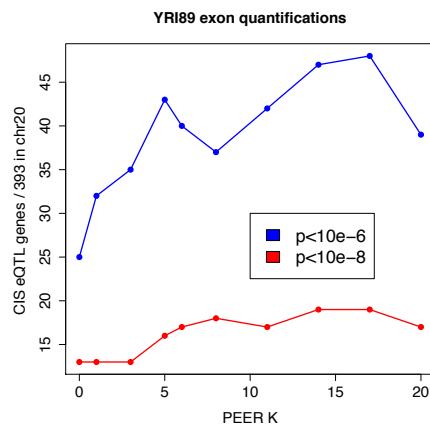
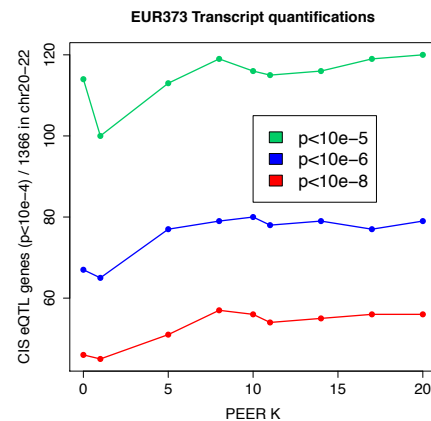
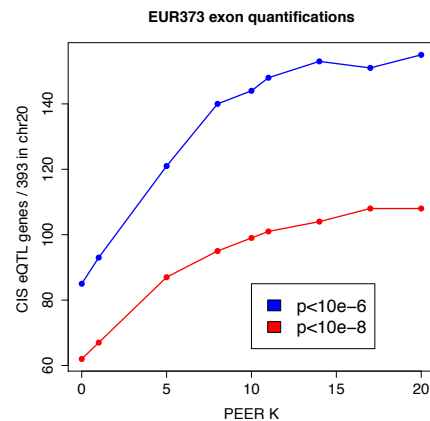


Expression data normalization

- Exon quantifications
 - read count per exon normalized by the total number of mapped reads
 - exons that have >0 quantification in >50% of the individuals (across the whole dataset)
- Reducing technical variation
 - correction for sample processing covariates and QC covariates is difficult
 - analysis has been done and is informative of QC aspects, but very difficult and subjective to pick covariates to correct for
 - PEER - factor analysis to find synthetic covariates from expression data
 - better power and less subjective
 - separately for EUR & YRI quantifications, with sequencing lab and population (for Europeans) as additional known covariates that are corrected
 - residuals of this analysis = corrected quantifications
 - additional transformation to standard normal distribution for eQTL analysis with linear model

PEER expression normalization

- empirical testing which number of covariates gives the highest power
 - EUR373 exon: K=15
 - Eur373 transcript: K=5
 - YRI89 exon: K=5



eQTL analysis

- Linear regression model with the Matrix-eQTL tool
 - $\text{PEER_corrected_normalized_expression} \sim \text{genotype} + \text{genotype PCs}$
 - cis-eQTLs 1 MB on both sides of transcription start site
 - 5000 permutations of for each exon quantifications -> individual p-value limit for each exon
 - the exact same analysis for EUR373 transcript quantifications just to demonstrate that it's better to use exons

Further analysis

- Same analysis for miRNA quantifications
- Splicing QTLs – run as a separate analysis (maybe just for chr 1-3) to test what's the best phenotype to use
 - Exon link ratios (Altrans algorithm by Halit Ongen from UNIGE)
 - Exon junction ratios
 - Exon junction quantifications corrected for the expression levels of both of the exons
 - Exon inclusion percentage (from Pedro)
 - similar methods as for eQTLs: PEER, transform to standard normal
- Targeted trans-analysis
 - cis-eQTLs
 - miRNA variants and miRNA eQTLs
 - coding variants
- Transcriptome QTLs : joint analysis of all transcriptome phenotypes and multiple independent QTLs per gene. The idea:
 - all phenotypes per gene – exons, exon junctions, N-TAR quantifications, gene fusions – normalized and transformed to standard normal
 - 1st round cis-QTL analysis for all phenotypes -> get the best variant-phenotype association per gene (if significant = under permutation p-value threshold)
 - 2nd round cis-QTL analysis for all phenotypes correcting for the variant and phenotype of the 1st QTL -> get the best SNP-phenotype association if significant
 - 3rd round cis-QTL analysis for all phenotypes correcting for the 1st and 2nd QTL -> get the best SNP-phenotype association if significant
 - continue as long as significant associations are found